Hieu Tran

Implementation of high integrity file system for serial flash memories

Subtitle

Metropolia University of Applied Sciences

Bachelor

Information Technology

Thesis

Date

|  |  |  |
| --- | --- | --- |
| Author(s)  Title | | Hieu Tran  Implementation of high integrity file system for serial flash memories |
| Number of Pages  Date | | xx pages + x appendices  5 May 2010 |
| Degree | | Bachelor |
| Degree Programme | | Information Technology |
| Specialisation option | | Smart Systems |
| Instructor(s) | | Keijo Länsikunnas, Title (for example: Project Manager)  First name Last name, Title (for example: Principal Lecturer) |
|  | | |
| Keywords |  | |

**Contents**

1

[1 Background 1](#_Toc23227041)

[1.1 What is file system 1](#_Toc23227042)

[1.2 File and Directory 1](#_Toc23227043)

[1.3 Basic File System Operations 3](#_Toc23227044)

[1.3.1 Initialization: 3](#_Toc23227045)

[1.3.2 Mounting 3](#_Toc23227046)

[1.3.3 Unmounting 4](#_Toc23227047)

[1.3.4 Files manipulation (creating, opening, writing, reading, deleting, moving, renaming) 4](#_Toc23227048)

[1.3.5 Directories manipulation (creating, opening, writing, reading, deleting, moving, renaming) 5](#_Toc23227049)

[1.4 Different type of existing FS 5](#_Toc23227050)

[1.4.1 Traditional FS (block based file system) (conventional) 5](#_Toc23227051)

[1.4.2 How do they store data in block device (inode, direct, indirect blocks) 5](#_Toc23227052)

[1.4.3 Bitmap 6](#_Toc23227053)

[1.4.4 I-node 6](#_Toc23227054)

[1.4.5 Core of an i-node: the data stream. 6](#_Toc23227055)

[1.4.6 Advantages 7](#_Toc23227056)

[1.4.7 Disadvantage 7](#_Toc23227057)

[1.5 Flash memory overview 8](#_Toc23227058)

[1.5.1 How does flash storage device works 8](#_Toc23227059)

[1.5.2 Flash limitation in hardware 9](#_Toc23227060)

[1.5.3 Managed flash vs. Unmanaged flash 10](#_Toc23227061)

[2 My FS 10](#_Toc23227062)

[2.1 SpinFS (abbr for SPI NOR File System) 10](#_Toc23227063)

[2.1.1 Storage format 10](#_Toc23227064)

[2.1.2 Node data structure 11](#_Toc23227065)

[2.1.3 How it works 13](#_Toc23227066)

[2.2 Garbage collection 15](#_Toc23227067)

[2.3 Operations example 16](#_Toc23227068)

[2.4 Integrate into FUSE lib 16](#_Toc23227069)

[3 Testing result 16](#_Toc23227070)

[3.1 Hardware in use 16](#_Toc23227071)

[4 Conclusion 17](#_Toc23227072)

[4.1 What design goal has been achieved 17](#_Toc23227073)

[4.2 Limitation of current design 17](#_Toc23227074)

[4.3 Further development 17](#_Toc23227075)

[4.4 What I achieve with this project 17](#_Toc23227076)

Appendices

Appendix 1. Title of the Appendix

Appendix 2. Title of the Appendix

Maybe have

Introduction

File system Background

Flash Technology Background

# Introduction

Early File system in computing, the need of file system

Distinct characteristics for each storage media and operating system

The rise of Flash storage device. Number of flash storage device <ref?>, flash market

# Background

## What is file system

File system is a piece of software that allows Operating System to manage data on permanent storages. These permanent storages, which often are optical disks, hard disks, flash storage drives, etc., are used to store computer data and information for long-term even without electrical power <ref?>. Therefore, the main function of file systems is providing a way to organize, store, retrieve, and manipulate data and information on such kind of storages. To bring the most natural experience for end users when storing and retrieving computer data to and from persistent storage devices, hierarchy concept has been adapted to the computer world with two main components: file and directory. Thus, file systems are usually built around what is called file and directory.

## File and Directory

From human world, or library in particular, when storing books with careful thought about ease for later look up and retrieval they should be put into shelves sorted by family name of the authors or categorized by their subjects. File system can be compared to a library where a file is a book that has individual piece of information and directory is a shelf that holds multiple books or files that somehow related to each other. This relationship between file and directory makes organizing and looking for computer data more modular.

From file system point of view, a file has some information that is used by the computer, and in order to call the exact piece of data that stored in a file, a name must be assigned to the file. As a result, the job of file systems is storing files, which includes the raw data or "stream of bytes" and its names, on the storage device, then should be able to return back the content of the file when computer requests by passing its name.

Most computer users are already familiar with functionality of a file or a directory from direct experience when using the computers. Computer programs write data permanently to storage drive into files, which can be simply seen as a piece of information. A bit of information can be some text, graphical image, data structure, any combination of data types or arbitrary bytes only the program can interpret.

"A file appears as a continuous stream of bytes at higher levels, but the blocks that contain the file data may not be contiguous on disk." <ref1 p.11> File systems is responsible for mapping logical memory in file to physical memory on disk, thus when there is a write request, correct physical blocks will be modified and data on corresponding blocks needs to be returned in case of read request.

Directory (or Folder in Windows terms) is the way file system organize multiple files. Directory is basically a file, but its content is a list of names, these names can be name of files or names of sub-directories (directories living under a directory). How does this list of names store in storage plays a crucial part in file lookup performance? If this list is unsorted, searching for a file name in the list must scan through the whole list. Normal file systems take advantage of sorted data structure when storing content of directory for efficiency in lookup operation.

File systems are usually packed within Operating Systems and each Operating System has its own implementation of file systems. Depending on the underlying physical storage medium, purpose when storing data and host Operating System, file system type will be chosen for individual media. For instance, CDs use ISO 9660 file system, hard disks and managed flash drive like SSD depend on what is available from the host OS, Microsoft Windows utilizes NTFS while MacOS takes advantage of Apple's APFS with improved GUI experience and Linux has ext4 developed by the Open-source Community <ref?>.

Terminologies related to file systems

(INTEGRITY libguide p.96 file system terms: file system implementation, file system instance, metadata, mount, partition, root file system)

Partitions: are divisions or parts of a real hard disk drive. A partition is really only a logical separation from the whole drive, but it appears as though the division creates multiple physical drives, which can be managed separately. A partition is a logically independent section of a hard disk drive that contains a single type of filesystem.

Volume: refer to a disk or partition that has been initialized with a file system.

File system instance (or filesystem): refer to all the files, directories and file system data structures stored in a partition of a device.

Metadata: is data information that provides knowledge and description about other data. For example, creation time of a file is a valuable information about a file, but it is not part of the data stored in the file.

Logical Volume (#TODO to add)

Superblock: a small area on a volume where characteristics of the filesystem currently residing on that volume are stored. The information a superblock usually contains are volume size, block size, layout and counts of empty and filled blocks, with other file system specific metadata such as size and location of inode tables, disk block map and usage information.

Inode: is a data structure of a file system implementation (usually on Unix-like operating systems) where all the information about a file except its name and its actual data are stored. The i-node also provides the connection to the physical locations on disk having the file's data.

File system vs filesystem

## Basic File System Operations

### Initialization:

First operation to attach file system to any fresh partition in a drive is always creating an empty file system on a given volume. The volume's characteristics like total size and also user's preferences will be taken into account when creating and placing internal data structures of file system onto the fresh volume. Most of the time, this related to the size of superblock and how does files and directories will be stored later on.

During the process of initializing file system, empty top-level directory must be created, also known as root directory. Every files and directories created later will be organized under this root directory in the hierarchy. In Linux world, this is usually done by executing mkfs programs, whereas Windows users are more familiar with this process under the name Formatting in Formatting a disk or USB drive.

### Mounting

Mounting is the task of accessing the raw device, using file system program to read its superblock compute necessary information for later reading and writing its content. Even mounting is often a user program, it will return a handle to Operating System for accessing the disk. In graphical OS environment, mounting is transparent and performed automatically when there is a new storage device connecting to the computer. One significant job when mounting a disk is to check the consistency state of the disk, was it cleanly shutdown last time, or there is a power outtage while data was being written to the drive. Such inconsistency can cause corruption to the disk and all data will be lost. In case, file system detects unclean shutdown state of the disk when trying to mount, extra task of verifying and possibly repairing any damage must take action. This task will be handled by another program called file system check, or fsck for short, and it is extremely complex and takes time since whole storage needs to be checked.

### Unmounting

The umount operation involves flushing all unwritten data from RAM to physical media, marking the volume to be "clean" indicating that a normal shutdown is performed and removing the accessing handle from Operating System. After unmounting, it should not be possible to access the storage device until the next mount operation.

### Files manipulation (creating, opening, writing, reading, deleting, moving, renaming)

After mounting a fresh file system, there is nothing on the volume until a new file is created and some data is stored in the file to be written permanently to disk. Creating a file requires name of the file and which directory the file will reside in. If the volume is empty, the new file will be created under root directory. File creation operation only involves allocating inode and writing metadata of the file, the actual data, stream of bytes in the file will be added later when writing to it actual happens. Since having a new file under a directory means the content of the directory has been modified, file system also needs to update the parent directory's content.

When a file is presented on file system, it needs to be opened before being able to read and write. Similar to mount file system, opening a file returns file handle to Operating System and by using this handle, Operating System can request file system to read or write to a specific location in the file.

Write operation allows computer programs to store data while read allows extracting information in files. File system needs a reference to the file, which is returned from open operation, offset position in the file to begin reading and writing, memory buffer and length of the requested writing or reading data. Writing to a file can increase file size and in such case, extra memory blocks will be allocated to append to the end of logical file memory array, which eventually update the file i-node and potentially superblock of the file system depending on its implementation. Writing to a logical address of a file will be translated to a physical address by file system, thus making it permanent. Read operation is much simpler, all file system needs to do is map logical position of the file to the corresponding physical block on disk, which comes down to retrieving data from storage device and place it to user's buffer.

Beside creating file, it is only natural to have option to delete file. File deletion process first removes the name of the file in parent directory's name list and until there are no more programs with open file handle to the file, file system can free the file's resources by returning occupied blocks to the free block pool and the file i-node to the free i-node list. It is not necessary to set all the bit associated with the metadata and data deleted file to be zeros, marking a delete flag in the file metadata and free these block in used/free block pool is enough.

Rename is often regarded as the most complex operation that a file system has to support. There are many checkings and validations must be passed before actual renaming takes place. These checks include whether the source and destination names are in different directories, then new name can be the same as the old one, if names refer to directories, new name cannot be sub directory of the old name, etc. Only after all validations are satisfied, file system will delete old name entry, create new name entry in new destination directory then update corresponding parent directories.

### Directories manipulation (creating, opening, writing, reading, deleting, moving, renaming)

As mentioned above, directory is basical a file where its data is the name list of sub files and directories. Therefore, most file operations can be applied to directory.

In summary, file system is a program that enable Operating System to store, retrieve and manipulate data on persistent storage device, data here is usually stored as content of files and organized into directories. Each file system targets a specific type of storage on a specific Operating System.

## Different type of existing FS

### Traditional FS (block based file system) (conventional)

All started by FFS, the grandfather of most modern file system, set the standard for robustness and speed <ref1 p-33>

Each file system has its own design and implementation that serve a specific purpose with targeted hardwares including host systems and the storage devices it will be used on. However, the design below tend to be generic and provide adaquate performance with many features for user-friendliness and security. Most of the idea about superblock, block bitnmap, i-n ode bitmap, direct, indirect-blocks has been used in file systems such as BSD FFS, BFS, and FAT.

### How do they store data in block device (inode, direct, indirect blocks)

Essentially, there is a known place on a storage device to store the superblock which is the most important block to the file system. Suitable choices for this place are often the first or the last block of the partition but first block are usually chosen for the ease of seeking it (no need to find the size of the partition to calculate the last block number). In this superblock, there are block size, total number of blocks, number of used blocks, dirty bit flag, and address of i-node of the root directory. Withouth this reference from superblock to the root of the hierarchy of all files and directories, the file system would have no way to find any files on the volume.

### Bitmap

Following blocks after the superblock can be occupied for bitmap scheme, which is an approach to managing free space on a disk. The bitmap scheme represent each disk block as 1 bit, thus binary value of 0 or 1 in a bit can indicate vacancy status of a block whether it contains invalid data that the file system can freely write to it or it is not emptied and should not be overwritten. The number of blocks are used for bitmapping is totally based on the size of the partition and file system block size. Each byte consists of 8 bits, therefore the bitmap for 8GB disk with 1K blocks would requires 1MB of space, or 1024 blocks in this case.

<equation>

### I-node

Beside metainformation about the file such as the size of a file, access permission information, its creation and modificaton times, i-node data structure needs to keep track of which locations on disk are belongs to this i-node data stream. This basic structure is the fundamental building block of how data is stored in a file on a file system.

### Core of an i-node: the data stream.

Traditional approach for linking on disk addresses to logical file offsets is storing a list of blocks directly inside i-node, which is called direct blocks. Each entry in this list is a physical block address of the storage device, and since the size of an i-node structure is limited, it limits the amount of data the file can contain. Generally, about 4 to 16 block addresses can be stored directly in i-node, which means maximum size of a file can be is 16KB with 1KB file system block size. To address this space constraint, indirect block can be used. Rather than having direct reference to a physical address that has the file data, i-node can carry block address of this indirectly block. While data block contains user data, indirect block has pointers to other data blocks that do have user data in them, which make up the whole stream bytes of file data when combining together. Therefore, one disk block address can map a much larger number of data blocks, instead of mapping 1 by 1 with direct block address.

<figure?>

Indirect blocks help increasing the maximum data size in a file an i-node can keep track of, however, it is not enough to locate the data blocks of a file much more than a few hundred kilobytes in size. To overcome this issue to allow an even bigger file, indirect block technique can be applied a second time, making double-indirect blocks. The same concept and basic idea still hold true for double-indirect block as indirect block. Each double-indirect block address that i-node contains points to a block on disk whose content is more pointers to indirect blocks and repectively refers to exponentially amount of actual data blocks constructing a file.

<figure?>

### Advantages

robust, straight-forword easy to implement

Traditional design of file systems is straight-forward, there are not many abtraction layers between how users see files and directories on computer screen and how file system hide its implementation details to store those data on disk. Basically, data in a file is broken down into many chunks of blocks and file i-node is responsible for keeping track of location of those blocks. On top of that, file system superblock will always have connection to the address where that i-node resides and return that information whenever there is a request to read/write that file. Subsequently, when accessing a file, in addition to disk operations of accessing directly the blocks having actual file data, there will be few more reads to superblock and other blocks to search for the i-node and eventually leading to the data blocks. These additional reads are called file system overhead, and because of the simplicity in the design, traditional file systems performance suffer little from it and tend to be robust since less complexity means less bugs and corruptions. When putting more optimization features such as bigger such as block size and caching, performance can even be pushed further.

### Disadvantage

Writing any new data to the disk can require updating file system housekeeping data structure, which are superblock and bitmap, and having these areas to get frequently updated is a big disadvantage in this design. It really depends on the underlying storage device, continually changing the content in a part of a hard disk drive might not be a problem, but doing the same thing in modern flash drive is not at all recommended due to the drive's characteristics and operations involving in writing data to a memory block which will be discussed in the following section about "Flash storage".

One key disadvantage of flash memory is that it can only endure a relatively small number of write cycles in a specific block.

## Flash memory overview

#TODO many plagiarism

Flash memory is an electronic non-volatile computer memory storage medium that can be electrically erased and reprogrammed. Flash storage device is the successor of hard disk drive, which use mechanical movement to read and write memory <todo>.

Non-volatile storage technology that does not require power to retain data

Was invented by Toshiba in 1980 based on EEPROM technology

Toshiba introduced NAND Flash for the first time in late 80s

Individual flash memory cell consisting of a FET transistor and floating gate

Floating gate is used to store cell's value

No electrons on the floating gate -> cell is in the erased ("1") state and has a low "turn on" threshold voltage (Vth)

Electrons on the floating gate increases Vth voltage -> Cell is in programmed ("0") state

Write operation:

A voltage aplied to the control gate causes a tunnel current to flow through the oxide layer, thereby injecting electrons into the floating gate

Erase operation:

A voltage applied to the silicon substrate releases the electrons accumulated at the floating gate

Flash gets its name from the requirement to do all bit bulk flash erase of the whole sector, block to bit 1

Flash memory is a modern type of computer memory constructed from semiconductor components and used primarily in persistent storage mediums. This memory type falls into category of non-volatile technology meaning that power is not required to retain data in the memory making it a perfect match to use as secondary storage, or long-term persistent storage.

It was invented by Fujio Masuoka <ref?> while he is working at Toshiba in 1980 and made commercially first introduction to the market in late 1980s <ref?>. The key component that composes individual flash cell is floating-gate MOSFET (metal-oxide-semiconductor field-effect transistor), also known as a floating-gate transistor. Differentiate from a normal MOSFET <pic?>, a floating-gate MOSFET has an extra gate called floating gate added between the control gate and the body of the transistor <pic?>. The floating gate is separated from the control gate and the body by the oxide insulating layers preventing any electrons on floating gate to escape easily. Having this electrical isolated element allows charges on floating gate to stay for long periods of time, and flash memory utilizes this characteristic to store cell’s value on floating gate. Theoretically, if there is no electron on the floating gate, the cell is in erased state and the bit value is known to be 1, whereas having electrons on floating gate presents bit value of 0.

NOR flash 1984

NAND flash 1987

### How does flash storage device works

each cell > floating trnasistor > 2 charge states > 1 binary (1 when no charge, 0 when) > SLC > more than 2 charge state > more than 1 bit encoded per floating gate > MLC more than 1 bit per cell > smaller tolerance (increase Vth voltage) > reading > negative charge on floating gate screen off some positiv charge on control gate > need more charge to reach threshold > current vs gate to source voltage plot > aplly intermidiate voltage in between 2 threshold voltage and measure the current > same thing apply for mlc flash + plot > writing (moving charges to and from the floating gate) > program (inject electrons into the floating gate '1') > erase (release electron accumulated at the floating gate '0') > 2 methods (not go into details) > limitation

Flash memory stores information in an array of memory cells made from floating-gate transistors. each cell stores only one bit of information in single-lvel cell (SLC) devices, while multi-level cell (MLC) devices, TCL, QLC, PLC can store more than 1 bit per cell.

Flash memory is based on floating gate transistors, which is a variant of MOSFET with one small change. There is an extra gate that is added between the control gate and the body, this extra gate is known as the floating gate is electrically isolated. It had no electrical contacts which means that any charges put on the floating gate will stay there for a long time (for years in fact, and that's what allow flash storage to store data without the need of power source.

A flash memory device has all the data encoded as a bunch of charges on these floating gates.

This section will discuss more in details of how read and write operations in flash storage device are performed with principles down to flash memory cell level. As introduced in previous section, each memory cell consists of a floating transistor which can have two charge states on the floating gate, either no charge citing bit value of 1 and negative charge implying bit value of 0, and this type of two states cell called SLC (single-level cell) is capable of storing one bit of data. However, to achieve bigger bit density in a single flash IC <abbr>, MLC (multi-level cell) technology was developed to allow more than one bit to be encoded per floating gate, resulting in increasing maximum capacity of flash memory significantly. The theory behind MLC is based on the fact that the number of electrons which can be charged into the floating gate is a variable. Therefore, instead of having only charge and no charge state, floating gate can analogously have as many charging states as it wants depending on the amount electrons presented on the floating gate. Even though technology always seeks to have the greatest number of states as possible, it comes with a downside when evaluating the state of floating gates. More charge states express smaller tolerance in distinguishing them leading to higher chance of errors both in determining the amount of charge and injecting correct number of electrons. Considering the size of a typical flash cell is only ??nm <ref?>, it will need to be extremely precise when measuring any aspects of it. Until today, most common types of MLC flash devices have 2 to 4 bits per cell, which requires 4 to 16 states of the floating gate to be distinguished.

<fig?> more bits means smaller reliability margin

Reading the value of a cell is basically detecting which charge state the floating gate is having, and from that bit value can be translate. Based on the principle operation of MOSFET that if there is a threshold voltage applied to the gate, current will flow from source to drain <ref? physic semiconductor book>; and by applying some voltage to the control gate of the floating-gate MOSFET, the existent of current from source to drain can disclose the charge state of the floating gate. The effect of floating gate in controlling the current is electrons on the gate act as a negative mask layer which screen off some positive charges on the control gate, and the strength of this influence depends on the amount of electrons there is. Therefore, the voltage required for the control gate when the bit value is 0 is higher than when the bit value is 1 (no electrons on the floating gate) in order to have current flows in the drain terminal. In <fig?>, the threshold voltage for erased state is called VT1 and threshold voltage for programmed state is called VT0; and according to the statement above, VT0 > VT1.

<fig?> demonstrates the relationship between the current flowing from source to drain versus gate to source voltage in an IV (Current vs Voltage) graph with two “turning on” curves starting at VT1 and VT0, corresponding to two threshold voltages of different charge states. An intermediate voltage in the middle of VT1 and VT0 indicating by the blue line on <fig?> can be applied to measure the current. When identifying the charge state on floating gate to determine the bit value of a flash cell, no reading on current value means the transistor is on “0” curve correlating to bit value 0, while positive current reading means the other curve and bit value is 1.

The approach for MLC flash memory is not too different, instead of two curves account for two threshold voltage levels, there are four curves, one for each charge state, present in the IV graph <fig?> if each cell encodes 2 bits of data. On this condition, there are three intermediate voltages and by reading the current value when applying each one voltage difference from gate to source, bit values can be identified.

Beside reading, the other fundamental operation on a flash cell is writing, which is the process of alternating number of elections on floating gate to reach different charge states. Despite no electrical contact to the floating gate <ref?>, few methods of sending charges through the oxide layer has been discovered. First method is Quantum Tunneling and the second one is Hot Electron Injection. This document is not addressing these subjects in details; however, in short Quantum Tunneling is based on quantum physics to shrink the oxide gap between control gate and floating gate to push electrons off from control gate to floating gate and vice versa <fig?> <ref?>, whereas, Hot Electron Injection increases kinetic energy of electrons from body substrate to let them hop over the insulating layer onto floating gate <fig?> <ref?>.

### Flash limitation in hardware

Wearing

Equal distribution, load balance

Both writing methods (quantum tuneling and hot injection) involve high voltages and high electric fields and this limits the number of times you can write to the floating gate transistor. What happen is that the electrons gain a lot of energy and dissipate that energy by colliding with the oxide layer lattice and this damage builds up over time. Oxide insulator wears out over time causing electron leakage that alters floating gate charge. Defined Vth limits are no more accurate and value detected wrongly. Once you've done enough writes, the damage is great enought taht the device becomes unusable.

SLC 100,000 writes, MLC 1000 - 10,000. That's small is because of the lower tolerance is involded for mlc so it takes less wear to make MLC unsuable.

Because of the limited number of writes flash flash memory needs to have some sort of wear leveling.

If the wear leveling wasn't there you might end up writing to the same area to the same block of memory over and over again and make that block of memory unusable very quickly, so we're leveling everns out the load so you don't get one region of memory wearing out quickly.

Don’t get wearing out much quicker than the rest of the flash

Reduce capacity of the flash

Flash technology is gaining its popularity and slowly becoming the successor to replace hard disk drives for secondary storage due to its faster read and write speed since there is no mechanical part movement involve like in the latter <ref?>, though one major limitation it is having is flash memory cannot stand the test of time. This issue of the technology is called memory wear, since flash memory has a finite number of program – erase (P/E) cycles before it becomes unreliable for reading. The reason behind memory wear connects directly with the current ways how the memory cell is written. Both writing methods (Quantum Tunneling and Hot Electron Injection) require high voltages and high electric fields to be applied on the transistor, which causes the electrons to gain huge energy then dissipate that energy by colliding with the oxide layer lattice <ref?>. This damage builds up overtime and oxide insulator wears out eventually, causing electrons leakage on the floating gate, thus charges get altered unintendedly and bit value is encoded wrongly.

On MLC flash memory, wear problem has even greater effect due to the small tolerance of the reliability margin <fig?>. Illustrating in <fig? error in Fresh MLC>, threshold voltage region of each charge state is much smaller than in SLC memory and potentially crosses over others when worn out, represented by dotted grey line, hence tiny error in number of electrons can cause incorrect reading when applying the same factory defined intermediate VTHR voltages for measuring. Apparently, due to the smaller tolerance involved for MLC, it takes less wear to make MLC flash unusable. Typical maximum number of writes before wearing happens for SLC is 100,000 P/E cycles while this number for MLC only ranges from 1,000 to 10,000 cycles <ref?>.

The consequences from wearing effect only rise when some blocks in the memory has significant higher erased count than the rest of the blocks because their data were constantly focused for rewriting. The heat map in <fig?> show an example for such situation where about 20% random locations in the flash is reaching the maximum P/E cycles and other flash blocks still has considerably low erased counts. In such condition, once those higher rewritten rated blocks wear out, the whole filesystem residing on this flash memory is acknowledged to be corrupted and the flash is not meant for continued use, even though 80% of the flash is still in good shape. Therefore, it is not a good use case of flash memory to get frequently updated and rewritten partially, this will decrease the flash’s life span tremendously. In order to extend practical used time of flash memory, data written to the storage should be distributed equally to all blocks, thus having erased counts across the flash to be closed to each other. <fig?> portrays what a healthy flash memory array should look like when all blocks have approximately similar erased counts.

To guarantee written data is distributed equally across the whole flash, at least one type of wear levelling should be present in the whole system. Wear levelling solution can be accomplished in form of hardware as an extra IC in the same package of the flash or executed as software directly from the host computer. In summary, if no wear levelling was integrated in a system that use flash storages to even out the load, applications might end up writing to the same block of the memory over and over again, thus making that block of memory unusable quickly.

### Managed flash vs. Unmanaged flash

Modern flash storages can be divided to two types, managed flash and unmanaged flash, based on the existence of a component called controller in the same storage package.

Managed flash

Popular memory devices such as SSD, SD, eMMC, UFS found in PCs, laptops, mobile phones nowadays are all managed flashes. There is a controller on the device that do many memory management applications specifically for the current type of flash such as ECC error correction, wear leveling, bad block management, etc. By having a dedicated controller for the flash in the same package, all management operations are handled by the memory internally, lifting the burden from host computer to take extra care for any of these services. Moreover, integrated controller has better understanding of the underlying structure and characteristics of the flash than any external controller can possibly support, which proves internal controller is likely to have better optimizations at doing any mentioned operations.

Unmanaged flash

Raw NAND/NOR chips, mostly used for embeded devices with limited resources.

Unmanaged flashes are also known as raw NAND/NOR chips, which essentially consists of only memory array ICs without any kind of dedicated memory management component. This type of storage tends to have smaller footprint and lower cost comparing to its managed counterpart, thus suitable for embedded applications where they are used in an often limited resource systems.

Since there is no onboard memory management, all responsibilities to provide these instructions rely on host side. Therefore, when implementing file system targeting to use for such devices, wear leveling needs to be taken into consideration somehow that data will be written evenly to all blocks of the flash. If the design of the file system is comprehended with traditional approach, first few blocks of the flash where superblock and bitmap are resided will have larger P/E cycles than any other blocks.

Below graph shows an example when using FAT file system (which uses some beginning blocks for storing file system metadata, i-node table in this case) directly on a raw flash without wear leveling, which should always be avoided.

<pic>

To sum up, traditional file systems are inadvisable choices to manage files on to of raw flashes because of obvious reason, they are wear levelling intolerant. Therefore, new file system designs need to be invented to conform with the natural characteristics of flash memory.

# My FS

## SpinFS (abbr for SPI NOR File System)

### Storage format

spinfs is a log-structured file system

Differentiate from traditional block based file system that storage location is bound to a piece of data, log-structured file system make use of the entire storage for a circular log which is appended with every change made to the filesystem.

Unlike their block based coutnerpart, which has i-node blocks representing a file or directory scatter around the whole disk having an address pointer pointing to the blocks that have the actual data which also scatter around, this circular structure consists of multiple nodes, each representing a file or directory and the data inside it.

### Node data structure

The structure starts with a common header containing the i-node number of that node and all metainformation for that i-node, and may also carry a variable amount of data.

In spinfs, file and directory are treated equally. Directory is theoretically a file, whose data is a list of file names which are the name of the files and sub-directories are under this parent directory. Moreover, both of them share some similarities in their natural information. For example, both file and directory must have name, i-node number referring to it, its creation, modification time, size of it, and who is the owner and who can access. From these matching statistics of file and directory, spinfs is considered to have only one raw\_node data structure that can be used for both of them, which minimize the complexity of its design.

The spinfs-node structure is

<node structure>

Spinfs node is very similar to conventional i-node that it contains the metainformation about entities that live in the file system with only one exception that instead of storing addresses of that entity's data blocks which live somewhere else, data part is included directly inside spinfs node structure right after metadata part.

Looking at raw\_node structure, it is clear to see some basic file information spinfs supports. First of all, maximum file name length is 32 one-byte ASCII characters, which is said to be the minimum in an interactive system <p18. ref1>.

The next field, inode\_num, is self-explain, it is the i-node number of file or directory this raw\_node is referring to. In spinfs, 32-bit unsigned integer value is used for storing i-node number, allowing for 4 milliard files to be existed on the filesystem. However, considering that raw flashs are usually appeared in embeded systems, this maximum amount of files are unlikely to be reached due to the simplicity and often single, straight forward application these kind of systems provide <ref to avarage number of files in a filesystem>. One thing to noted that i-node numbers are never reused in spinfs. In case when files are deleted, new files will always have the next highhest available i-node number rather than having an obsoleted number from deleted files.

The uid/gid fields record ownership information about a file. Spinfs is designed to be run from on Linux host machines, thus it follows convention that was specified in POSIX that any file must have corresponding user id and group id which this file belongs to <ref to POSIX>. Combining with mode field, the file system can provide file access permission check. By following POSIX specification, file permission model in spinfs consists of user, group and "other" classes and three distinct operations; read, write and execute; these classes can do to a file system entity. The checking is done by comparing current logged in user with uid and gid fields to determine which class file access permission will be checked, then mode field will tell if the user can continue doing what he/she intends to do with the file, either read, write to it or execute an executable file. In addition to file access permission, spinfs also stores information about whether this raw\_node is a regular file or a directory along in this 32-bit value. (INTEGRITY libguide p.96 file system terms)

flags (obsolete, deleted)

Moving on to the next flags field, it is a record of various bits of i-node states, and at the moment there is only deleted state beside normal state of an i-node in spinfs.

Spinfs maintains some useful timely aspect of files known as creation time and last modified time of a file, and with these information users can easily query for files with a particular timestamp in mind. Unlike others Unix file systems, spinfs does not attempt to support last accessed time simply because this information is too expensive to maintain (every access to a file will need to update its node structure, and in a log-structured file system like spinfs, this means the whole new node with all file content will be written to the flash even each time a small part of the file is read, which eventually wears the flash extremely fast) over the small amount of use that it gets.

The parent\_inode field is effectively the i-node number of the parent directory of this file/directory. With parent i-node stored directly inside a node structure, traversing backward the filesystem tree can be made efficiently to reconstructing a full path name of the file, while without it, the only way to know the full path name of an i-node is having that information in memory all the time while the file is opened.

The version field is a important information for any nodes because it is where spinfs maintains historical ordering for each i-node. As the spirit of spinfs is a log-structured file system, every update to a file is actually writing a new file to storage file rather than modifying old data blocks, the file system needs a way to identify which is the most recent version of a file among many nodes referring to the same file. It is exactly what this field in raw\_node structure is responsible for, each new node is written with a version higher than all previous node belonging to the same i-node. Similar to inode\_num, version is an 32-bit unsigned integer, hence there is a ceiling limit for a number of nodes to be written for each node during the whole flash chip's lifetime, however, this amount of time is limited so this limitation is deemed to be acceptable.

The data\_size field indicates how many bytes is the size of raw\_node data part, which essentially is the size of a file in case this node is a regular file node.

The last field of this data structure is a flexible array member, data[], where the actual data stream of a file is located. Since files' size vary, length of data[] array is determined by data\_size field, thus the starting address of the data is always presented in the node structure while the last address can be easily calculated for the file system to know where is the next consecutive node.

After some considerations and updates to the design of spinfs, the author decided to implement checksum for each node to improve integrity of the file system. In C programming, flexible array member has to be the last data member of a struct <C99 ref> so a new field to store checksum cannot be added to the end of the current data structure, thus leaving the author of spinfs no choice other than having it in data[] array right after the actual data of the node. Consequently, this increases the size of each node by the size of the checksum, ?? bytes for SHA1 sum in this case, at the end and data\_size is no longer reflect the actual data size of a file if raw\_node is a regular file node but the actual size plus ?? bytes.

By always writing to the end of the flash in circular motion, all blocks in the flash will be written at the same rate and have even number of erase counts

### How it works

mkfs.spinfs

In order to use any file system with a storage device the very first operation is initializing the medium with a format that file system can understand. According to spinfs, this step is essentially erasing the whole raw flash device and write the first node for root directory, which is also the most important entity in any file system. Writing root directory can be as simple as transfering all bytes consisting in a raw\_node structure with "/" name, i-node number of 1, parent i-node of 0, since root directory has no parent, and data part has 0 byte size to address 0x0 of the flash.

create

The next usual operation with file system is populating the storage by creating files or sub-directories under root directory. New file created in the filesystem means a new node structure referring to the file with corresponding i-node will be written to the flash, continue after the last occupied address of the last node, in the case with a freshly formatted filesystem, new node is added after root directory node. Continue creating new files and sub-directories will keep appending new nodes for those entities to the flash toward the ending address. At the point when there is not enough space to allocate new node, an operation called garbage collection will be triggered trying to search for free space, and if it fails to do so, a "Not enough space left" error will be returned. Garbage collection will be discussed more in details later in this section.

update

Due to the design of log-structured file systems, every update to content or metadata of a file will result in writing a whole new node for that file with modified data to next available space in the raw flash. The new node will have higher version value than the old ones, and these old nodes are said to be obsoleted, where the content they contain has been outdated by a later node. Space taken by obsoleted nodes is referred to as "dirty space" and will be reclaimed later by garbage collection operation.

delete

File deletion is also considered an update to the file. Spinfs simply marks the file as deleted by setting deleted bit in its flag and write a new version of the file's node with zero data to the flash. After the file is marked as deleted, any operations with associated i-node number should not be possible and return error.

note about file/dir creation

When creating or deleting any files and directories, content of the parent directory is subjected to get modified, hence an update to the parent's node will be performed and another new node will be written.

read

need to find the latest version of the node

For the reason that there will be many nodes associated with a single i-node number in file system structure, when reading content of an i-node, spinfs should ignore all obsoleted nodes and only return the latest information about that i-node. Therefore, spinfs always keeps an i-node table in memory and updates this table with corresponding changes. The i-node table has columns for i-node number, latest version of raw\_node for that i-node and physical starting address of that raw\_node in storage device. At mount time, spinfs scans through the whole flash to populate this table. Upon looking at a new i-node number, a new entry will be appended to the table, while upon a newer version of an entry is found, that entry will be updated with the corresponding data.

## Garbage collection

head and tail store in Security Register, int32, so can store # slots,

At the point when the system is out of space due to continuous writes of new data, it needs to start reclaiming the dirty space which is the result of obsoleted and deleted nodes. To assist this reclaiming operation, the file system always keeps track of physical offsets in flash media of the oldest node and the next erased address for a new node, these address offsets are called head and tail respectively. In a filesystem that this cleaning process has never been triggered, the head will stay at the very beginning of the flash, whereas the tail is reaching the last available address.

(Should the following paragraph be in Testing Hardware section)

During practical work of this project, the test flash chip model has feature called Security Registers, which are small separate memory regions aside from the main memory region consisting of only 256 Bytes, and two of these registers are chosen solely for the purpose of storing the head and tail pointers’ values. In spinfs, head and tail are 32-bit unsigned integer, thus each Security Register can store 64 values for them and will be updated according to each write to main memory region. This means any new node written to the end of the flash will also write a new value for both head and tail in Security Registers 1 and 2. When all 64 slots has been used, spinfs simply erased these two registers and start writing new head and tail values from the beginning. The search for correct, or most recent, head and tail values in Security Registers is straightforward since the registers can be read backward from the end and the first non-erased value, not 0xFFFFFFFF, is the wanted one.

Each time garbage collection code is executed, the objective is to erase the first flash block pointed by head. The operation will begin by iterating head toward tail in natural direction over each node and examine whether the pointing node is obsolete or valid. In case the node is obsoleted by a later version written with the same i-node number or it is marked with a deleted flag with zero sized data part, it will be skipped, and the head moves on to the next node.

On the other hand, when the node is still valid, meaning that it has the latest version of a still-in-use file, the whole node will be copied to the tail of the log by writing an exact same node just with bigger version number, hence rendering the current node to obsolete and the head continues iterating.

Once the head has been progressed to the next erased block boundary of the flash, all nodes in the previous block should not be relevant to the filesystem anymore and it is safe to erase that block without losing any data. As a result, the amount of free space in the flash, indicating by number of bytes counting from the tail address to the head address, increases by an erased block size. In instance of insufficient dirty space to have size of one erased block when added together, garbage collection process will return error.

## Operations example

## Integrate into FUSE lib

# Testing result

## Hardware in use

This project was developed on a Raspberry Pi and a NOR flash chip soldered to its breakout circuit board which then is connected to GPIO pins corresponding for SPI communication channel 0 on the Pi. The specific model of the Pi is Raspberry Pi 2B #FIXME, and the flash is NOR Flash IC S25FL164K from Spansion #FIXME. The Pi is acting as the host running Raspbian which is a distribution of Linux Operating Systems targeted to Raspberry Pi’s hardware. From the breakout board of the flash chip, jumper wires were used to connect directly to GPIO pins on the Pi as shown in <pic> and described in <table>. Because spinfs was written with C programming language, there is a need to interface SPI communication to the code, and the author has chosen to use SPI APIs from WiringPi library. For example, to send over a SPI packet, programmer can call wiringPiSPIDataRW(SPI\_CHANNEL, buf, sizeof(buf)) function. This library can easily be installed from apt-get package manager in Raspbian and should be linked with when compiling C source files with “-lwiringpi gcc” option.

The S25FL164K flash model is a 64 Megabits (8 Megabytes) variant in its flash chip family, thus having address counting from 0x000000 to 0x7FFFFF. All flashes in this family share some similar characteristics, such as 256 bytes programming page size and smallest erased size is 4 Kilobytes. It is a NOR flash meaning any byte in the flash are allowed to get accessed directly for reading and writing; comparing to NAND flash, this has to be performed on the whole page. As being said, a read command can read one byte at any specific address or data in following addresses will also be returned as long as clock line in SPI protocol is still running. However, program to the flash accept only 256 bytes at most <ref. 2>; and if this maximum value is desired to be written, the address used in the program command should be 0xXXXX00, otherwise the address will be wrap around at page size and old data will be overwritten.

# Conclusion

## What design goal has been achieved

Wear leveling

Data integrity with checksum, tail update after writing new node, so in case of power failure during write, nodes before tail are always valid (code challenge to verify this and clean up bad write)

## Limitation of current design

write to flash every update

Not power cut tolerance

## Further development

## What I achieve with this project

References

Details of the references are given here. Use the referencing system required in your degree programme or as agreed with your supervisor.

Layout of this page in the author-date /Harvard) referencing system:

Details of the reference Details of the reference Details of the reference Details of the reference Details of the reference.

Details of the reference Details of the reference Details of the reference Details of the reference Details of the reference.

Layout of this page in the number (Vancouver) referencing system:

1. Details of the reference Details of the reference Details of the reference Details of the reference Details of the reference.
2. Details of the reference Details of the reference Details of the reference Details of the reference Details of the reference.

Title of the Appendix

Content of the appendix is placed here.

Title of the Appendix

Content of the appendix is placed here.